3 · 4 LUGLIO 2024

Data quality framework per coniugare la qualità delle statistiche ufficiali con le innovazioni della statistica, del data science e dell'intelligenza artificiale

MAURIZIO VICHI

Professore Ordinario, membro ESGAB | Sapienza Università di Roma,

CNS15ª

Conferenza Nazionale di **Statistica**

La statistica ufficiale nel tempo dell'Intelligenza Artificiale

#CNStatistica15

Indice

How: come la Statistica sta cambiando anche con il Data Science e l'Intelligenza Artificiale

Trasformazioni significative dovute ai progressi tecnologici e all'aumento esponenziale dei dati in termine di volume, varietà, velocità

1 **Tendenza**: -> **dalle Indagini Statistiche face to face all'uso osservazioni dirette di eventi rilevanti e registrati elettronicamente, eventualmente integrati da indagini (esempio lampante è il censimento permanente)**

Dataficazione della Società e produzione di Dati Organici (dati generati automaticamente)

- Incremento Comunicazione Mediata dal Computer (CMC) e dell'Internet of Things;
- Piattaforme di social media che trasformano gli aspetti della vita in dati quantificabili ("dati organici");
- Approfondimenti in tempo reale per molti fenomeni rilevanti.

2. **Tendenza: -> da piccoli a grandi campioni di dati**

integrando Dati di indagine + Dati amministrativi + Big Data

- Uso di diverse fonti di Dati di indagine, dati amministrativi e Big Data (dati detenuti privatamente e dati IoT);
- Descrizione più realistica dei fenomeni mediante molte variabili (esempio: andare oltre il PIL);
- Uso di serie storiche multivariate e di dati spaziali con grande granularità **per spiegare fenomeni complessi nel tempo e nello spazio**

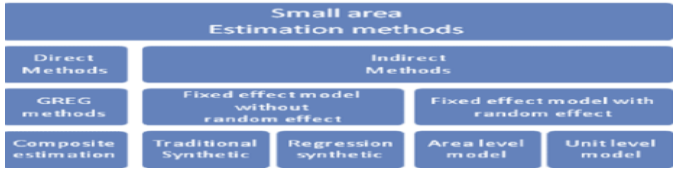
3. **Tendenza**: -> **da statistiche univariate/bivariate a statistiche multivariate ad alta intensità di computer + Intelligenza Artificiale**

Integrazione di metodologie di statistica con apprendimento (superv. e non) + Data Science + AI

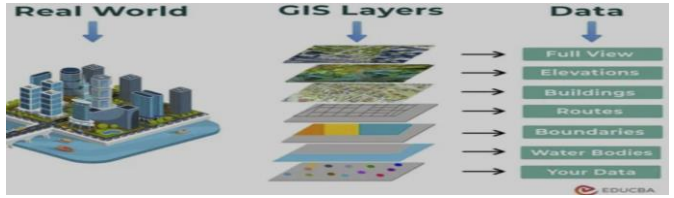
- statistica multivariata e analisi dei dati con inferenza (ricampionamento);
- Statistica inferenziale ad alta intensità di computer (simulazione, ricampionamento);
- Intelligenza artificiale con modelli statistici spiegabili (-> XAI) IA spiegabile;

Which and Why: Innovazioni metodologiche che hanno un grande impatto sulla statistica ufficiale

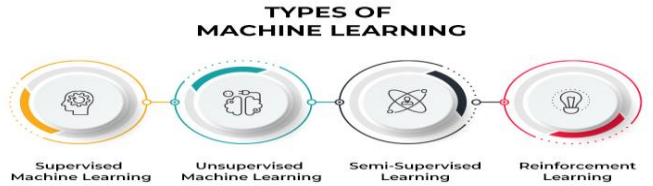
- **Data Linking, Integration, and Big Data:** *Integrating datasets* with, record linkage, and data fusion enhances the scope and quality of statistical outputs;
- **Computer-Intensive Inferential Statistics:** Techniques like Monte Carlo Simulation; Bootstrap, and Markov Chain Monte Carlo help to *reflect uncertainty in complex conditions*.



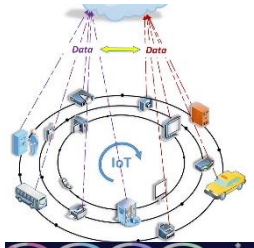
- **Small Area Estimation (SAE):** This method allows for *reliable granular estimates for small geographic areas* or subpopulations, like estimating poverty rates at a district level;



- **Geospatial Statistics:** Incorporating Geographic Information Systems (GIS) aids in assessing spatial relationships, crucial for mapping and analyzing phenomena;



- **Machine Learning and AI:** Advanced algorithms for classification and regression improve the accuracy of predictive models, and explainable AI ensures transparency in statistical predictions;



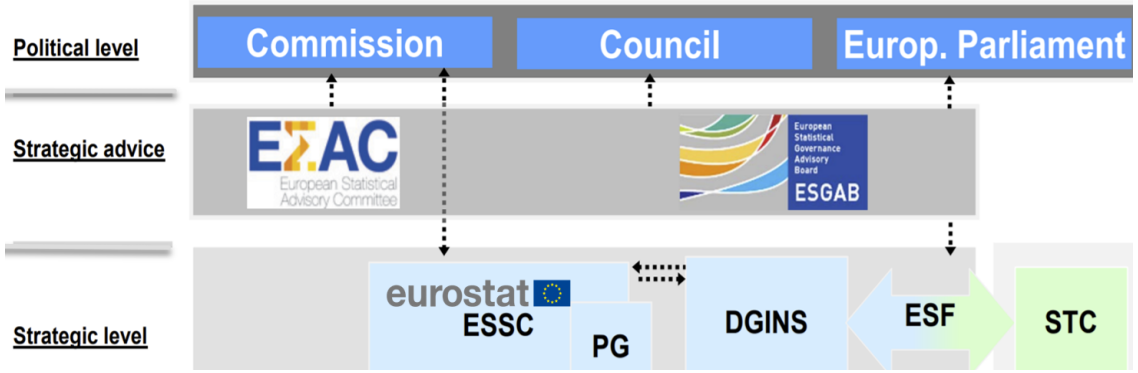
- **Real-Time Data Collection and Crowdsourcing:** Technologies such as web *scraping* and *high-frequency IoT data collection* provide real-time insights, valuable for monitoring rapidly changing phenomena. Crowdsourcing and citizen science experiments.

- **Data Visualization and Communication:** Advanced visualization tools make statistical information more accessible and understandable, aiding in the dissemination of complex data.

When: Priorità dei cambiamenti della Governance Statistica necessari per le innovazioni

- La legge 223 europea e ancor più la legge italiana 322/del 1989 sono un modello di governance statistico lontano dalle innovazioni e che non sempre riescono a rispondere appieno alle attività di produzione della Statistica moderna e a tutte le innovazioni introdotte da Internet e dagli avanzamenti tecnologici

Quale modello di governance per il Sistema Statistico Europeo all'altezza delle Innovazioni Statistiche? Ne stiamo parlando in Europe in ESAC e ESGAB e vorrei provare a fare alcune considerazioni su come l'impostazione Europea si può adattare al Sistan al fine di avere una Governace coerente con il sistema Europeo



Si devono distinguere tre livelli per la governance
Politico: Presidenza del Consiglio, Parlamento (producono l'impianto legislativo)

Strategico Consultivo: rappresentato dalla COGIS con doppia funzione (completezza e qualità dell'informazione)

Livello Strategico ISTAT -> COMSTAT -> SISTAN

Per il livello strategico a livello europeo si prefigura un **European Statistical Data Ecosystem** con un modello di **governance adattivo**.

Che cosa vuol dire Governance Adattiva?

1. Il quadro giuridico e normativo consente flessibilità e risposta rapida alle esigenze emergenti in materia di dati e ai progressi tecnologici;

2. La Governance adattiva migliorare la resilienza e la capacità di risposta dell'Ecosistema alle sfide emergenti e impreviste, come pandemie guerre e crisi economiche;
3. E' un approccio più flessibile e dinamico alla governance, che evolve le tradizionali procedure associate al Programma Statistico Annuale. e promuovere l'agilità, la flessibilità e le capacità di risposta rapida all'interno del sistema statistico.

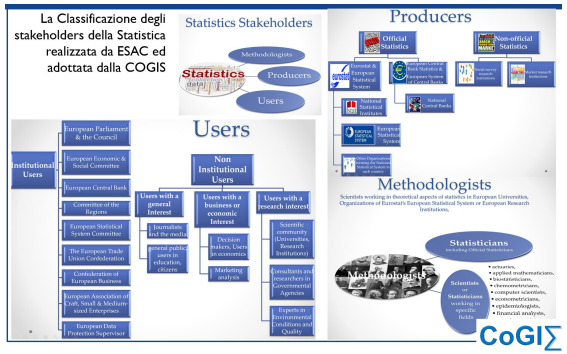
What : Governance Statistica dei Dati

La GOVERNACE DEI DATI STATISTICI

- Nel contesto della governance dei dati statistici, la priorità più rilevante è lo sviluppo di ecosistemi di dati integrati con framework per gestire l'integrazione delle tre fonti di dati: dati di indagine, dati amministrativi e big data.

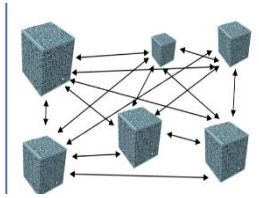
Il MODELLO

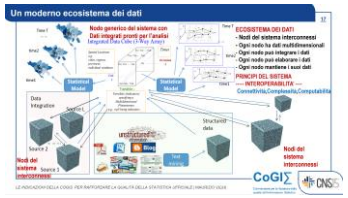
1. Il modello di ecosistema statistico richiede innanzitutto un apprendimento continuo dei fabbisogni di statistica da parte di un ampio spettro di utilizzatori sia quelli istituzionali (governo, parlamento, enti sul territorio,…) che quelli non istituzionali (ricercatori, giornalisti, …). La COGIS quale istituzione indipendente deve monitorare i fabbisogni di informazione, analogamente a ESAC (European Statistical Advisory Committee) e monitorare la qualità dell'informazione statistica, analogamente a ESGAB (European Statistical Governance Body);



L'ARCHITETTURA

2. L'architettura dell'ecosistema dovrebbe essere decentralizzata, ovvero formata dal network di istituzioni, circa 60, che contribuiscono al programma statistico. Ogni nodo gestisce i propri dati garantendo al contempo un accesso sicuro e controllato per la condivisione e l'integrazione dei dati;

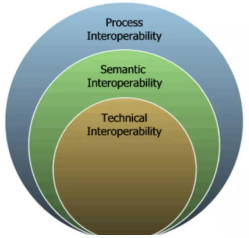




L'INTEROPARABILITA' e STANDARDIZZAZIONE

3. La progettazione dell'ecosistema richiede lo sviluppo e l'implementazione di standard e protocolli di interoperabilità completi per facilitare lo scambio continuo di dati tra i nodi. Ciò include l'accordo su formati di dati comuni, standard di metadati e protocolli di comunicazione e la creazione di API standardizzate e gateway di dati sicuri per facilitare l'integrazione e lo scambio di dati in tempo;

Interoperability and Standards



IL CLOUD COMPUTING e DISTRIBUTED COMPUTING

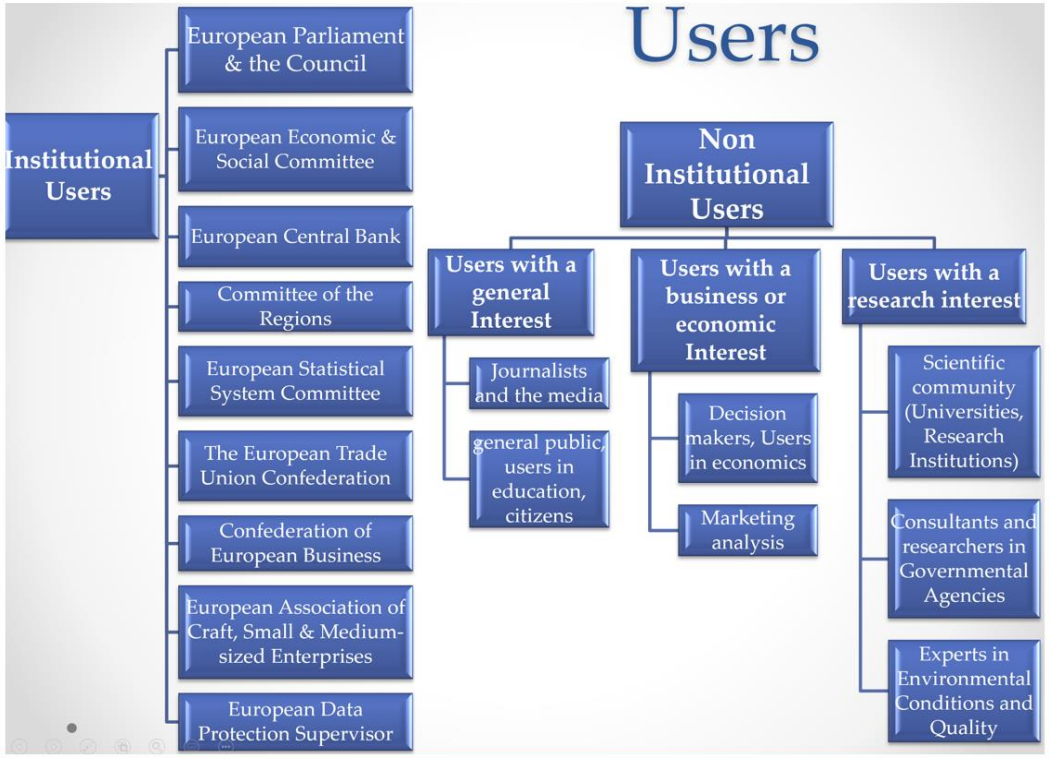
4. L'ecosistema dovrebbe utilizzare il calcolo distribuito e adoperare strumenti basati su cloud-computing per le funzionalità di trattamento e analisi dati garantendo in questo modo la standardizzazione del trattamento, analisi dei dati e produzione delle statistiche, sebbene i repository centrali vengano evitati per motivi di sicurezza e privacy;

What (1/4): Apprendimento continuo del fabbisogno di informazione

IL MODELLO di ECOSISTEMA

1. Il modello di ecosistema statistico richiede un apprendimento continuo dei fabbisogni di statistica da parte di un ampio spettro di utilizzatori sia quelli istituzionali (governo, parlamento, enti sul territorio,…) che quelli non istituzionali (ricercatori, giornalisti, …).
   La COGIS quale istituzione indipendente deve monitorare i fabbisogni di informazione, come ESAC (European Statistical Advisory Committee) e monitorare la qualità dell'informazione statistica dell'Ecosistema, come ESGAB (European Statistical Governance Body)

La Classificazione degli stakeholders della Statistica realizzata da ESAC ed adottata dalla COGIS



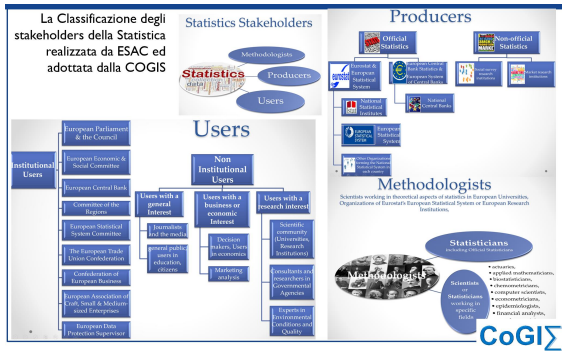
What : Governance Statistica dei Dati

La GOVERNACE DEI DATI STATISTICI

- Nel contesto della governance dei dati statistici, la priorità più rilevante è lo sviluppo di ecosistemi di dati integrati con framework per gestire l'integrazione delle tre fonti di dati: dati di indagine, dati amministrativi e big data.

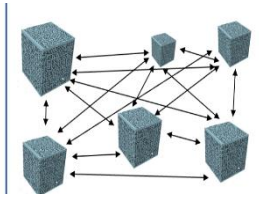
Il MODELLO

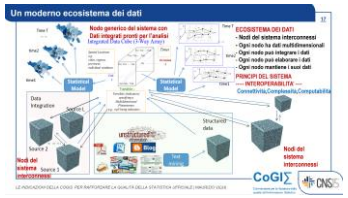
1. Il modello di ecosistema statistico richiede innanzitutto un apprendimento continuo dei fabbisogni di statistica da parte di un ampio spettro di utilizzatori sia quelli istituzionali (governo, parlamento, enti sul territorio,…) che quelli non istituzionali (ricercatori, giornalisti, …). La COGIS quale istituzione indipendente deve monitorare i fabbisogni di informazione, analogamente a ESAC (European Statistical Advisory Committee) e monitorare la qualità dell'informazione statistica, analogamente a ESGAB (European Statistical Governance Body);



L'ARCHITETTURA

2. L'architettura dell'ecosistema dovrebbe essere decentralizzata, ovvero formata dal network di istituzioni, circa 60, che contribuiscono al programma statistico. Ogni nodo gestisce i propri dati garantendo al contempo un accesso sicuro e controllato per la condivisione e l'integrazione dei dati;

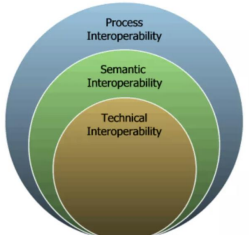




L'INTEROPARABILITA' e STANDARDIZZAZIONE

3. La progettazione dell'ecosistema richiede lo sviluppo e l'implementazione di standard e protocolli di interoperabilità completi per facilitare lo scambio continuo di dati tra i nodi. Ciò include l'accordo su formati di dati comuni, standard di metadati e protocolli di comunicazione e la creazione di API standardizzate e gateway di dati sicuri per facilitare l'integrazione e lo scambio di dati in tempo;

Interoperability and Standards



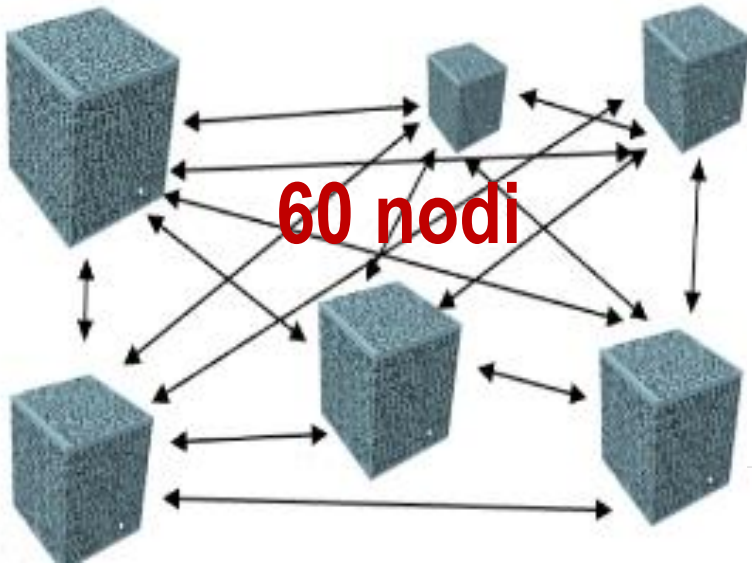
IL CLOUD COMPUTING e DISTRIBUTED COMPUTING

4. L'ecosistema dovrebbe utilizzare il calcolo distribuito e adoperare strumenti basati su cloud-computing per le funzionalità di trattamento e analisi dati garantendo in questo modo la standardizzazione del trattamento, analisi dei dati e produzione delle statistiche, sebbene i repository centrali vengano evitati per motivi di sicurezza e privacy;

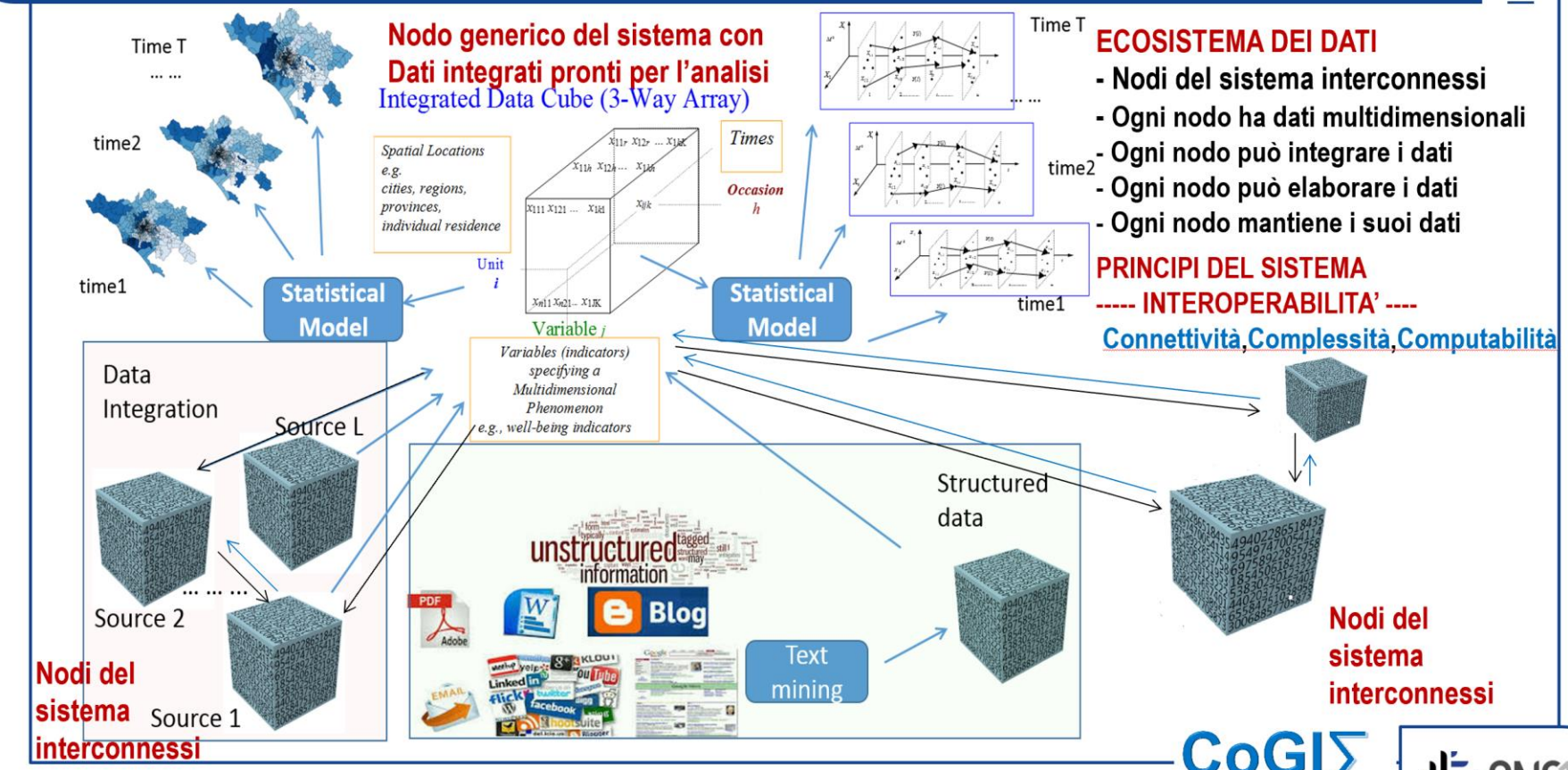
What (2/4): L'Architettura dell'Ecosistema dei dati

L'ARCHITETTURA dell'ECOSISTEMA

2. L'architettura dell'ecosistema dovrebbe essere decentralizzata, ovvero formata dal network di istituzioni, circa 60, che contribuiscono al programma statistico. Ogni nodo gestisce i propri dati garantendo al contempo un accesso sicuro e controllato per la condivisione e l'integrazione dei dati;



Un moderno ecosistema dei dati



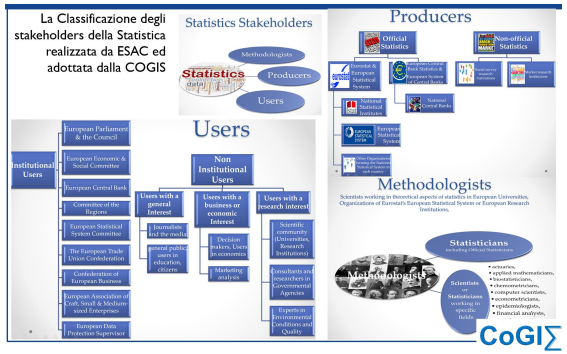
What : Governance Statistica dei Dati

La GOVERNACE DEI DATI STATISTICI

- Nel contesto della governance dei dati statistici, la priorità più rilevante è lo sviluppo di ecosistemi di dati integrati con framework per gestire l'integrazione delle tre fonti di dati: dati di indagine, dati amministrativi e big data.

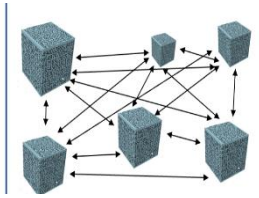
Il MODELLO

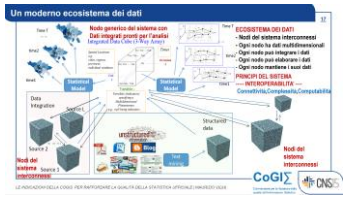
1. Il modello di ecosistema statistico richiede innanzitutto un apprendimento continuo dei fabbisogni di statistica da parte di un ampio spettro di utilizzatori sia quelli istituzionali (governo, parlamento, enti sul territorio,…) che quelli non istituzionali (ricercatori, giornalisti, …). La COGIS quale istituzione indipendente deve monitorare i fabbisogni di informazione, analogamente a ESAC (European Statistical Advisory Committee) e monitorare la qualità dell'informazione statistica, analogamente a ESGAB (European Statistical Governance Body);



L'ARCHITETTURA

2. L'architettura dell'ecosistema dovrebbe essere decentralizzata, ovvero formata dal network di istituzioni, circa 60, che contribuiscono al programma statistico. Ogni nodo gestisce i propri dati garantendo al contempo un accesso sicuro e controllato per la condivisione e l'integrazione dei dati;

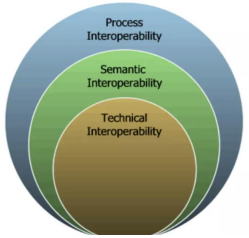




L'INTEROPARABILITA' e STANDARDIZZAZIONE

3. La progettazione dell'ecosistema richiede lo sviluppo e l'implementazione di standard e protocolli di interoperabilità completi per facilitare lo scambio continuo di dati tra i nodi. Ciò include l'accordo su formati di dati comuni, standard di metadati e protocolli di comunicazione e la creazione di API standardizzate e gateway di dati sicuri per facilitare l'integrazione e lo scambio di dati in tempo;

Interoperability and Standards



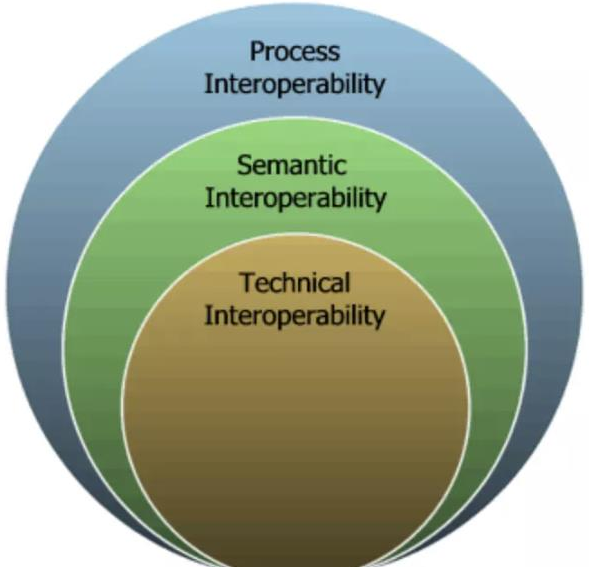
IL CLOUD COMPUTING e DISTRIBUTED COMPUTING

4. L'ecosistema dovrebbe utilizzare il calcolo distribuito e adoperare strumenti basati su cloud-computing per le funzionalità di trattamento e analisi dati garantendo in questo modo la standardizzazione del trattamento, analisi dei dati e produzione delle statistiche, sebbene i repository centrali vengano evitati per motivi di sicurezza e privacy;

What : (3/4) L'INTEROPERABILITA' e la STANDARDIZZAZIONE

L'INTEROPARABILITA' e STANDARDIZZAZIONE

3. La progettazione dell'ecosistema richiede lo sviluppo e l'implementazione di standard e protocolli di interoperabilità completi per facilitare lo scambio continuo di dati tra i nodi. Ciò include l'accordo su formati di dati comuni, standard di metadati e protocolli di comunicazione e la creazione di API standardizzate e gateway di dati sicuri per facilitare l'integrazione e lo scambio di dati in tempo;



Process Interoperability
refers to the ability of different nodes (Institutions) to work together effectively exchanging the data, but also producing new data an new processes such as the statistical annual programme together. It ensures that the workflow and the procedural aspects align across different institutions;

Semantic Interoperability
refers to the ability of different systems to understand and interpret the data exchanged meaningfully. It goes beyond simply exchanging data to ensuring that the receiving system can understand it in the same way the sending system intended;

Technical Interoperability
Refers to the ability of different systems to connect, communicate, and exchange data without considering its meaning. It covers the technical aspects like data formats, communication protocols, and interfaces.

Data Standardization
Format Consistency: Ensuring that data values are stored in a consistent format;
Terminology Harmonization: regularize terms and labels to use consistent nomenclature;
Data Types: Converting data into standard types;
Coding : Using standardized code sets for categorical data;
Removal of Redundancy and Data clining: Eliminating duplicate data and errors.

What : Governance Statistica dei Dati

La GOVERNACE DEI DATI STATISTICI

- Nel contesto della governance dei dati statistici, la priorità più rilevante è lo sviluppo di ecosistemi di dati integrati con framework per gestire l'integrazione delle tre fonti di dati: dati di indagine, dati amministrativi e big data.

Il MODELLO

1. Il modello di ecosistema statistico richiede innanzitutto un apprendimento continuo dei fabbisogni di statistica da parte di un ampio spettro di utilizzatori sia quelli istituzionali (governo, parlamento, enti sul territorio,…) che quelli non istituzionali (ricercatori, giornalisti, …). La COGIS quale istituzione indipendente deve monitorare i fabbisogni di informazione, analogamente a ESAC (European Statistical Advisory Committee) e monitorare la qualità dell'informazione statistica, analogamente a ESGAB (European Statistical Governance Body);

L'ARCHITETTURA

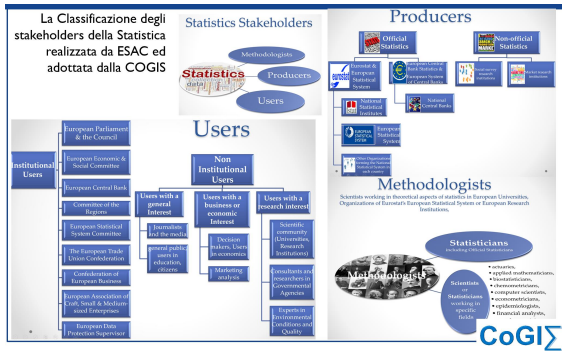
2. L'architettura dell'ecosistema dovrebbe essere decentralizzata, ovvero formata dal network di istituzioni, circa 60, che contribuiscono al programma statistico. Ogni nodo gestisce i propri dati garantendo al contempo un accesso sicuro e controllato per la condivisione e l'integrazione dei dati;

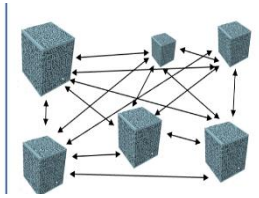
L'INTEROPARABILITA' e STANDARDIZZAZIONE

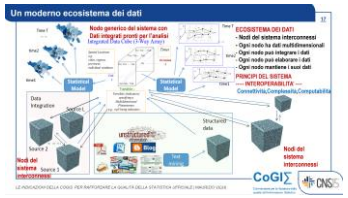
3. La progettazione dell'ecosistema richiede lo sviluppo e l'implementazione di standard e protocolli di interoperabilità completi per facilitare lo scambio continuo di dati tra i nodi. Ciò include l'accordo su formati di dati comuni, standard di metadati e protocolli di comunicazione e la creazione di API standardizzate e gateway di dati sicuri per facilitare l'integrazione e lo scambio di dati in tempo;

IL CLOUD COMPUTING e DISTRIBUTED COMPUTING

4. L'ecosistema dovrebbe utilizzare il calcolo distribuito e adoperare strumenti basati su cloud-computing per le funzionalità di trattamento e analisi dati garantendo in questo modo la standardizzazione del trattamento, analisi dei dati e produzione delle statistiche, sebbene i repository centrali vengano evitati per motivi di sicurezza e privacy;







What (4/4): IL CLOUD COMPUTING e DISTRIBUTED COMPUTING

IL CLOUD COMPUTING e DISTRIBUTING COMPUTING

4. L'ecosistema dovrebbe utilizzare il calcolo distribuito e adoperare strumenti basati su cloud-computing per le funzionalità di trattamento e analisi dati garantendo in questo modo la standardizzazione del trattamento, analisi dei dati e produzione delle statistiche, sebbene i repository centrali vengano evitati per motivi di sicurezza e privacy;

CLOUD COMPUTING

Scalability : Cloud computing allows for easy scaling of resources to accommodate varying data processing needs, which can be very beneficial for handling large volumes of statistical data;

Cost-Effectiveness: With cloud services, there's often a pay-as-you-go pricing model, which can reduce costs associated with maintaining and upgrading physical hardware;

Accessibility: Data and applications can be accessed from anywhere, facilitating collaboration among different institutions and stakeholders within the ecosystem;

Flexibility and Agility: Cloud services can be quickly adapted to new requirements without significant delays or costs, enabling quicker implementation of updates and new technologies;

Data Backup and Recovery: Cloud providers typically offer robust backup and disaster recovery solutions, ensuring data safety and resilience.

Vantaggio cruciale è la standardizzazione integrata dei processi. Tutti i nodi dell'ecosistema utilizzano gli stessi servizi e strumenti basati su cloud, c'è una naturale uniformità nei metodi di elaborazione e analisi dei dati impiegati.

DISTRIBUTED COMPUTING

Enhanced Processing Power: Distributed computing utilizes the collective processing power of multiple nodes, which can handle large-scale data analysis more efficiently.

Fault Tolerance: The decentralized nature of distributed computing provides better fault tolerance, as failure of one node does not lead to a complete system failure.

Improved Data Locality: Data can be processed closer to where it is stored or generated, reducing latency and providing faster insights.

Resource Optimization: Distributed computing can better utilize resources across different nodes, optimizing both performance and cost.

Vantaggio cruciale è la maggiore potenza di elaborazione. Distribuendo le attività di elaborazione dei dati tra più nodi, l'ecosistema può affrontare analisi statistiche complesse in modo più efficiente.

When: Il framework di qualità dell'informazione statistica. Il sistema di peer-review COGIS

- Il framework di qualità è basato sul **Codice delle statistiche europee** e la valutazione avviene con l'**European Peer Review**. Quest'ultima misura lo stato dell'arte del sistema. Come migliorare il framework di qualità?
- Serve un processo snello, ma formale di revisione peer review in cui le metodologie, le procedure di raccolta dei dati e i prodotti statistici siano regolarmente esaminati da esperti indipendenti della comunità scientifica;
- Sviluppo delle metriche con le istituzioni accademiche e di ricerca per costruire indicatori di qualità e le metriche di performance per garantire che soddisfino gli standard scientifici e professionali (*certificazione dei dati e dei processi*);
- Pubblicazione delle metriche di valutazione: dei risultati del processo di valutazione, mostrando come il feedback e l'input scientifico vengono utilizzati per migliorare la qualità statistica.

CoGIS

Commissione per la Garanzia della qualità dell'Informazione Statistica

Sistema di Valutazione Peer Review COGIS (VPRC): Visione Strategica:

Il sistema di VPRC è la strategia della COGIS per monitorare la qualità dell'informazione statistica. L'obiettivo è quello di esaminare la conformità e l'allineamento dei circa 60 Enti produttori di statistica (ONA e Uffici Statistici principali).

Obiettivo

- **Produrre la standardizzazione delle modalità di rilevazione, gestione, analisi, disseminazione e comunicazione delle informazioni statistiche**, sulla base del Codice Italiano ed Europeo delle statistiche ufficiali, **utili a definire Enti del Sistema statistico perfettamente integrati nel sistema, che possano cioè attingere e rilasciare facilmente informazioni**;
- **Identificare avanzamenti e progresso degli Enti del sistema** nell'allinearsi ai principi del Codice Italiano delle statistiche ufficiali;
- **definire raccomandazioni da indirizzare al SISTAN in generale per rafforzare la collaborazione tra Enti del sistema**;

Si propone un **approccio combinato** di **autovalutazione** e di **peer review con visita esterna** per trarre vantaggio dai due approcci.

1. La peer review è la modalità più diffusa nell'ambito della ricerca scientifica per valutare la qualità dei lavori pubblicati e creare un consenso scientifico, ovvero una opinione collettiva della comunità scientifica su un particolare argomento in una istante di tempo.

2. La peer review in questo caso ha il compito di indicare le aree di possibile miglioramento delle ONA e Uffici Statistici che producono informazioni del PSN nel loro complesso e di mettere in luce la presenza di buone pratiche in linea con quelle adottate a livello Europeo con il Codice Europeo delle Statistiche.

Audizione

La relazione finale del team di esperti valutatori (TEV) serve alla COGIS per l'audizione finale

3 · 4 LUGLIO 2024

grazie per l'attenzione

MAURIZIO VICHI

Professore Ordinario, membro ESGAB | Sapienza Università di Roma,

CNS15ª

Conferenza Nazionale di **Statistica**

La statistica ufficiale nel tempo dell'Intelligenza Artificiale

#CNStatistica15